International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 3, Issue 5, Dec 2013, 217-222
© TJPRC Pvt. Ltd.

EVALUATING AUDIO DESCRIPTORS FOR TIMBRE ANALYSIS IN SINGER IDENTIFICATION PROCESS

HARSHADA W. NIKAM

GHRCEM, Department of Computer Engineering, UoP, Pune, Maharashtra, India

ABSTRACT

This paper describes a method for singer identification based solely on the vocal contents of audio signal. The proposed method is evaluated on timbral features of any audio sample and then classifies feature vector using K-means algorithm and GMM(Gaussian Mixture Model).This work is focused on various audio descriptors of the timbre to analyze which features of the sound can be most useful for singer identification process. Basic and Spectral Audio descriptors like Spectral Centroid(SC), Spectral Roll-Off, Zero Crossing Spectral Kurtosis, RMS Energy, Shannon Entropy, Brightness, MFCC, Fundamental Frequency(f0) are considered. To evaluate the performance of the Spectral Features, various experiments are performed using text independent cohort GMM and K-means Clustering Algorithm.

KEYWORDS: Music Information Retrieval (MIR), Timbre Analysis, Singer Identification, Audio Signal Processing Application, Audio Feature Vector Generation

INTRODUCTION

Rapid growth in digital music database and increasing size of personnel music collection raised a demand of new retrieval methods, which are based on the contents of the music. One of the key content of any music sample is the vocal performance of any artist or singer. We can use these contents for Organizing, Browsing, visualising large music collection that also can be used in intelligent music recommendation and playlist generation systems.

A simple Structure of Singer Identification Systems is explained in following steps

- Vocal/Non-Vocal Segmentation.
- Feature Extraction.
- Classification.

Mentioned steps are well known and widely experimented in speaker recognition systems. In a proposed system the challenging part is a feature extraction where we have to represent Timbre in acoustic feature.

It has been observed that Timbre is not straightforward to define or measure, but some features are common in literature [1]. There exist an unbounded set of possible features one could extract from a audio signal. But we have to focus our attention on those that can be measured on an arbitrary frame of audio, so as to be able to characterise the instaneous timbre of voice sound.

LITERATURE SURVEY

The musical term timbre is used broadly to refer to the variability in sonic characteristics. It is generally considered conceptually separate from the aspects of pitch, loudness and duration, encompassing attributes which musicians might describe as “bright" vs. dull", “rough" vs. “smooth", etc. It is the attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and the pitch are dissimilar[2].Now, we can refer this timbre with number of audio descriptors. There are 70 different audio descriptors which are widely distributed. Objective of the experiments is to identify the minimal set of audio descriptors which are highly useful in the singer identification process and increases the accuracy of same.

Many toolkits are available today, which process the audio signal and generate the values for some of audio descriptors. We used MIR toolbox [3] for testing various audio descriptors and experiments the relativity of those to singer identification process .Audio Descriptors which are not included in MIR Toolbox but plays vital role in defining timbre of the sound were also included in experiments. These descriptors carry unique information about the recording. The audio descriptors can be one dimensional scalar values or a series of values resulting in feature vector.

Overall there seems to be a relationship in type of feature extractor used and the corresponding classifier used with given constraints on both, the input data file and the classifier.

Audio database is important in this whole process. The input audio files have various attributes such as file type (.wav, mp3), sampling rate, audio type (mono, stereo) bit rate etc. Some standard built in online databases are also available , but as this application is based solely on the vocal contents of the audio file it is required to generate a new audio database without any instrumental part and have same attributes for all audio files. It will help not only for generating effective feature vector but also increasing accuracy of the final result.

Following section describes the architectural model of the system.

SYSTEM DESIGN

A simple design is proposed for singer identification system, which includes following sections,

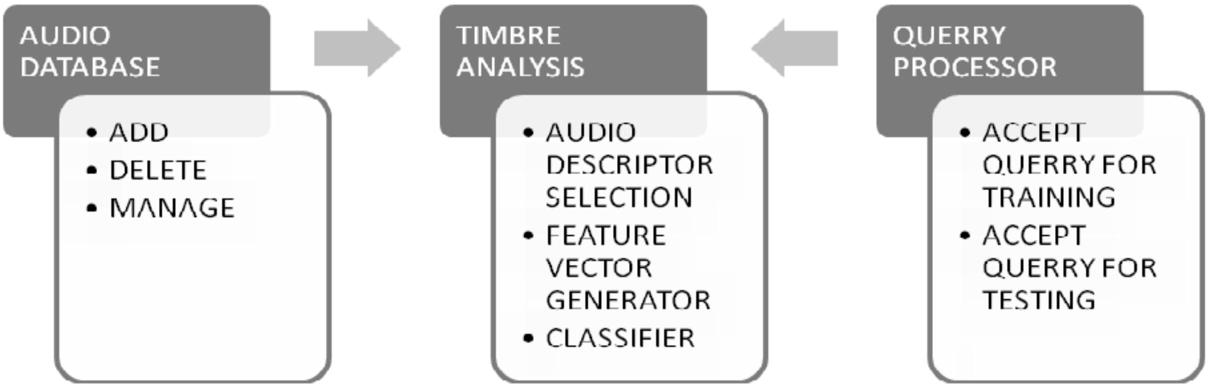


Figure 1: System Design

Audio Database is generated, initially it included 100 audio files from 10 different singers (10 each) sharing same file attributes like all are in .wav format, duration is 5 sec, sampled at 44.1 KHz, mono, bit rate is 128 Kbps. User can increase or decrease the size of database as per requirement. A GUI is provided to the user for selecting various timbre audio descriptors. Feature vector is generated according to the selection of the descriptors. This feature vector is then submitted to the classifier. User can also select classifier between K-means algorithm and GMM. User can train this timbre analysis module with various set of databases and test new unseen data. Selection process of audio descriptors as well as classifiers can be done dynamically, which gives wide scope for experimentation.

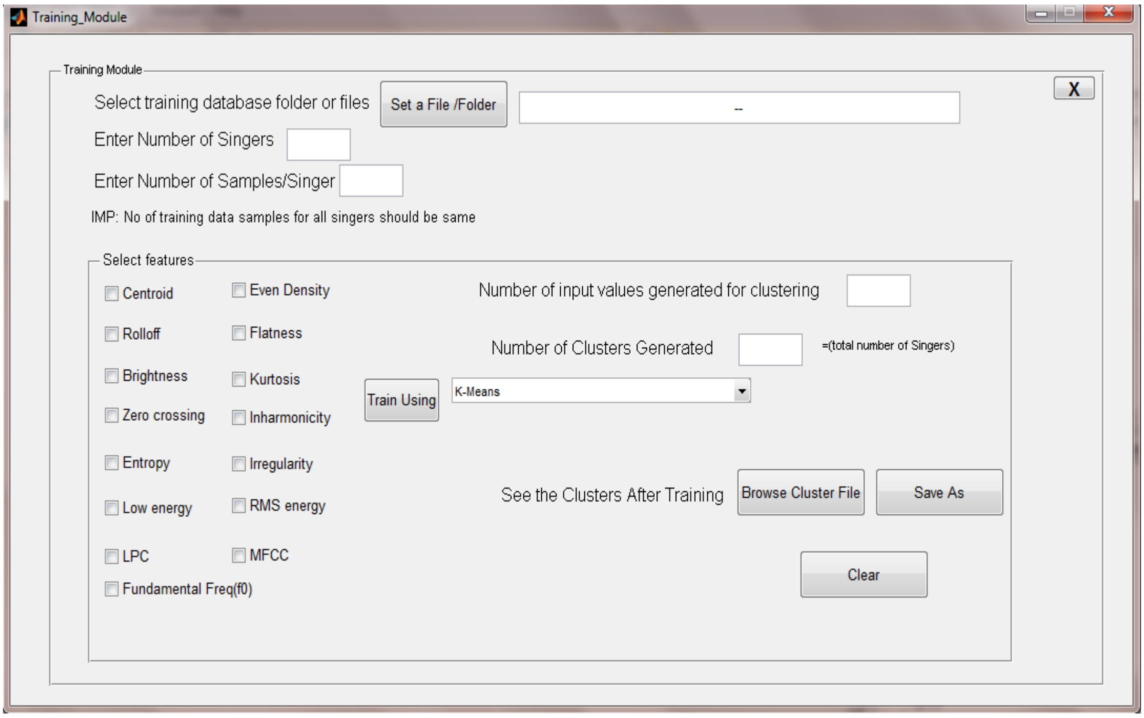


Figure 2: GUI for System

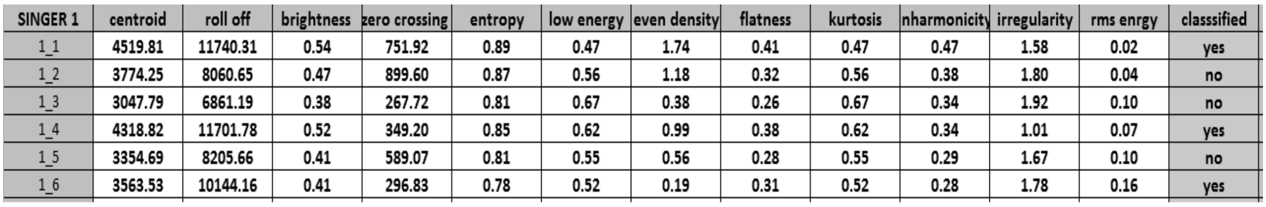
System focuses on audio descriptors. In experiments audio descriptors are selected by permutation –combination method .It generates feature vector in the form of matrix. Data is divided in to two parts for training and testing. Initially, 60% data used for training and 40% used for testing. In some next few iterations, Training data is reduced to 40% and rest of the data is used for testing.

K-means clustering algorithm handles vector values in very efficient manner but many audio descriptors like MFCC generate vector values. For fitting these vector values into feature vector matrix, mean value is taken. Accuracy of the result is then checked by analysing percentage of the accurate results generated by the system.

EXPERIMENTS AND RESULTS

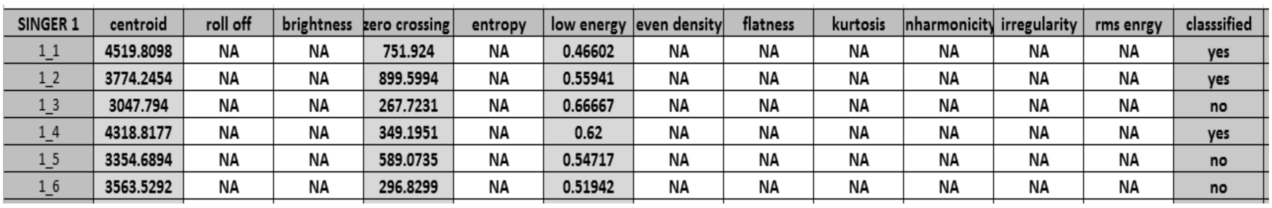
In first iteration, 60 audio samples (from 10 different singers-5 male,5 female, 6 each) were given to training module and all audio descriptors were selected. Generated trained cluster is saved in the database for further testing. In next iteration, audio sample remains same but audio descriptors are selected with the combination of two then three and so on. After every training process generated cluster is saved in same way in the database. Testing phase gives user ability to choose one of the saved cluster and test any unseen audio sample of any singer. Selection of audio descriptors and classifiers gives wide scope for experimentation, results are displayed as singer successfully identified if it was included in training phase or new singer identified.

Following table shows how generated results are stored and analysed, we can see initial training module database with all audio descriptors selected in figure 3 figure 4 shows successful classification with only two audio descriptor selected, whereas figure 5 shows successful classification with increased accuracy when six audio descriptor selected. The process is repeated and tested for all other singers sequentially and accuracy is checked over the size of training data and testing data with minimal selection of audio descriptors. Unlike K-means, percentage accuracy varies with singers in GMM which is only 38%. K-means gives successful results in 85% of cases in whole testing phase.



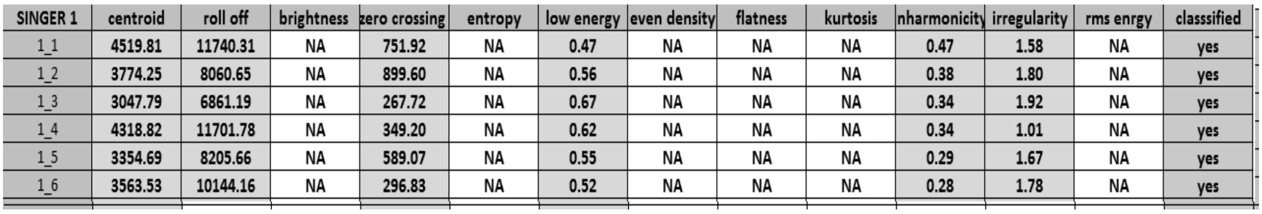
SINGER 1	centroid	roll off	brightness	zero crossing	entropy	low energy	even density	flatness	kurtosis	nharmonicity	irregularity	rms enrgy	classsified
1_1	4519.81	11740.31	0.54	751.92	0.89	0.47	1.74	0.41	0.47	0.47	1.58	0.02	yes
1_2	3774.25	8060.65	0.47	899.60	0.87	0.56	1.18	0.32	0.56	0.38	1.80	0.04	no
1_3	3047.79	6861.19	0.38	267.72	0.81	0.67	0.38	0.26	0.67	0.34	1.92	0.10	no
1_4	4318.82	11701.78	0.52	349.20	0.85	0.62	0.99	0.38	0.62	0.34	1.01	0.07	yes
1_5	3354.69	8205.66	0.41	589.07	0.81	0.55	0.56	0.28	0.55	0.29	1.67	0.10	no
1_6	3563.53	10144.16	0.41	296.83	0.78	0.52	0.19	0.31	0.52	0.28	1.78	0.16	yes

Figure 3: Snapshot of Database in Iteration 1



SINGER 1	centroid	roll off	brightness	zero crossing	entropy	low energy	even density	flatness	kurtosis	nharmonicity	irregularity	rms enrgy	classsified
1_1	4519.8098	NA	NA	751.924	NA	0.46602	NA	NA	NA	NA	NA	NA	yes
1_2	3774.2454	NA	NA	899.5994	NA	0.55941	NA	NA	NA	NA	NA	NA	yes
1_3	3047.794	NA	NA	267.7231	NA	0.66667	NA	NA	NA	NA	NA	NA	no
1_4	4318.8177	NA	NA	349.1951	NA	0.62	NA	NA	NA	NA	NA	NA	yes
1_5	3354.6894	NA	NA	589.0735	NA	0.54717	NA	NA	NA	NA	NA	NA	no
1_6	3563.5292	NA	NA	296.8299	NA	0.51942	NA	NA	NA	NA	NA	NA	no

Figure 4: Snapshot of Database in Iteration 2



SINGER 1	centroid	roll off	brightness	zero crossing	entropy	low energy	even density	flatness	kurtosis	nharmonicity	irregularity	rms enrgy	classsified
1_1	4519.81	11740.31	NA	751.92	NA	0.47	NA	NA	NA	0.47	1.58	NA	yes
1_2	3774.25	8060.65	NA	899.60	NA	0.56	NA	NA	NA	0.38	1.80	NA	yes
1_3	3047.79	6861.19	NA	267.72	NA	0.67	NA	NA	NA	0.34	1.92	NA	yes
1_4	4318.82	11701.78	NA	349.20	NA	0.62	NA	NA	NA	0.34	1.01	NA	yes
1_5	3354.69	8205.66	NA	589.07	NA	0.55	NA	NA	NA	0.29	1.67	NA	yes
1_6	3563.53	10144.16	NA	296.83	NA	0.52	NA	NA	NA	0.28	1.78	NA	yes

Figure 5: Snapshot of Database in Iteration 23

CONCLUSIONS

Basic Audio descriptors and spectral features can be use together for generating effective feature vector, which can explore the basic terminologies in Music Information Retrieval. At the end of this paper, we can say that combination of basic audio descriptors and spectral features improve the performance of MFCC based feature vector generation method and its use in singer identification system is very promising. we can also conclude that to get the accurate results we need to work really hard on standardized audio database, the effective fieldwork can help to produce effective result.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my Guide, course co-ordinator, colleagues in GHRCEM, as well as my family, friends and every other person who directly or indirectly involve in this project for being with me right through the study and providing valuable insights on the subject. Without your encouragement and guidance this work would not have existed.

REFERENCES

1. Dan Stowell, Making music through real-time voice timbre analysis: machine learning and timbral control, Queen Mary University of London ,2010.
2. ANSI. Acoustical Terminology. Number S1.1-1960. American National Standards Institute, New York, 1960.
3. MIRtoolbox 1.2.5 Olivier Lartillot Finnish Centre of Exce!ence in Interdisciplinary Music Research University of Jyväskylä,
4. Gaussian Mixture Model Classifiers Bertrand Scherrer February 5,
5. T. H. Park, "Towards Automatic Musical Instrument Timbre Recognition," November 2004.

6. G. Peeters, "A large set of audio features for sound description(similarity and classification) in the CUIDADO project," 23 April 2004.

7. R. H. Plomp, "Effect of Phase on the timbre of comples sounds," Journal of Acuostical Society of America, pp. 409-421, 1969.

8. M. S. S. G. a. K. J. Adams, "Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specificities, and Latent Subject Classes," PsychologicalResearch, vol. 58, pp. 177-192,

9. S. J. W. B. S. M. McAdams, "Discrimination of Musical Instruments Sounds Resynthesized with Simplified Spectrotemporal Parameters," JASA , vol. 2, p. 104, 1999.

10. W. Slawson, Sound Color, Berkeley: University of California Press, 1985.

11. "MPEG-7 Audio," MPEG, October 2005. [Online]. Available:

http://mpeg.chiariglione.org/standards/mpeg-7/audio .